# Analyzing the Scope of Conditions in Texts: A Discourse-Based Approach

Amanda Bouffier, Thierry Poibeau

# Analyzing the Scope of Conditions in Texts: A Discourse-Based Approach

**Amanda Bouffier* and Thierry Poibeau****

*TEMIS
164 rue de Rivoli – 75001 Paris
amanda.bouffier@temis.com

**Laboratoire d'Informatique de Paris-Nord
99, av. J.-B. Clément – 93430 Villetaneuse
thierry.poibeau@lipn.univ-paris13.fr

## Abstract

This paper focuses on the task that consists in automatically structuring free texts according to semantic principles, hence requiring a discourse analysis. We show that the task can be rephrased as a machine learning task in which the algorithm is supposed to take an optimal decision from the range of complex interacting constraints. The approach is implemented and evaluated taking the example of Health Practices Guidelines, medium-size documents intended to describe common practices that should be followed by physicians. Our approach outperforms previous approaches limited to sentence boundaries or requiring a lot of manual work.[1]

## 1 Introduction

The Semantic Web aims at producing knowledge from texts, which means normalizing and connecting pieces of information extracted from texts. Unfortunately, most of the time, the semantic blocks of information required by XML models are neither explicitly nor linguistically marked in the original text. These blocks of information correspond to discourse structures (Teufel, 1999).

In this paper, we addressed this problem for bio-medical texts. Health practice guidelines (HPG) describe best practices with the aim of guiding decisions and criteria in specific areas of healthcare, as defined by an authoritative examination of current evidence (Brownson et al., 2003). The Guideline Elements Model (GEM, `http://gem.med.yale.edu/`) is an XML-based guideline document model that can store and organize the heterogeneous information contained in practice guidelines (Shiffman et al., 2000). It is intended to facilitate the translation of HPG into a format that can be processed by computers. The main element of GEM, knowledge component, contains the most useful information, especially sequences of conditions and recommendations. However, calculting the scope of conditions remains a highly challenging task for NLP.

Even if some attempts have been done to analyse automatically the content of these documents (Shiffman et al., 2000; Georg and Jaulent, 2005; Bouffier and Poibeau, 2007), the analysis still require a huge amount of manual work. We propose here a new approach, based on simple machine learning techniques. We show that each relevant linguistic feature can be modelled as a constaint. Machine learning approaches allows one to determine automatically which features are the most useful ones for the task, as well as combinations of features. This approach yields slightly better results than previous ones requiring more manual work.

We first give a precise description of the task. We then present our discourse-based approach before giving details on the experiment we have carried out. Finally, we describe our results and discuss related work.

## 2 Automatically Filling the GEM DTD

Our main goal is to go from a textual document (see example 1) to a GEM based document, which structure basically corresponds to a tree (see example 2). We focus our analysis on conditions (including temporal restrictions) and recommendations, since these elements are essential for the task. Conditions and recommendations describe a complex network of nested elements that the parser must unravel.

The text of example 1 is complex and contains several levels of overlapping conditions. We observe a first opposition (*Chez le sujet non*

[1]This work is supported by Agence Nationale de la Recherche (ANR) through the TEXTCOOP and CROTAL projects.

⟦**Chez le sujet non immunodéprimé**⟧$_{cond1}$, ⟦**en cas d'aspect macroscopique normal de la muqueuse colique**⟧$_{cond2}$, [des biopsies coliques nombreuses et étagées sont recommandées (...)]$_{rec1}$. [Les biopsies isolées sont insuffisantes (...)]$_{rec2}$. [L'exploration de l'iléon terminal est également recommandée (grade C).]$_{rec3}$ ⟦**En cas d'aspect normal de la muqueuse iléale** (...)⟧$_{cond3}$, [la réalisation de biopsies n'est pas systématique (accord professionnel)]$_{rec4}$. ⟦**Chez le sujet immunodéprimé**⟧$_{cond4}$,[il est nécessaire de réaliser des biopsies systématiques (...)]$_{rec5}$

Figure 1: Example: extract of an HPG

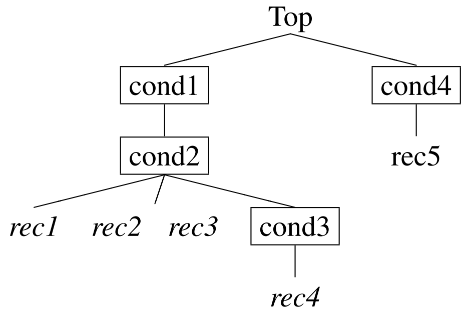


Figure 2: A tree reflecting the structure of the text presented in example 1.

*immunodéprimé/ chez le sujet immunodéprimé... Concerning the non-immunodepressed patient / Concerning the immunodepressed patient...*) but a second condition interferes in the scope of this first one (*En cas d'aspect normal de la muqueuse iléale... In case the ileal mucus seems normal...*). The task involves recognizing these various levels of conditions in the text and explicitly representing them through the GEM DTD.

Therefore, the problem mainly consists in discovering the scope of conditions. It can be rephrased as follows: given a recommendation, discover the set of conditions that rule its application. We detail our approach in the next section.

## 3 General approach

In this section, we first give an overview of our approach before detailing the set of constraints taken into account.

### 3.1 Input

The process takes in input a specific HPG. The document is automatically segmented into basic units (pairs of *governors* — conditions — and *targets* — conditions or recommendations), as described in (Bouffier and Poibeau, 2007).

Scope analysis is then performed, also based on a precise linguistic analysis of the content of representative HPG.

### 3.2 Extracting the Set of Relevant Lingusitic Features

A linguistic analysis of HPG led us to define the constraints which are relevant for the task. These constraints are heterogeneous. They include lexical items, co-referring cues, structural elements (elements reflecting the structure of the text), *etc.*

A precise analysis of relevant linguistic cues first has to be done (we just report here a brief summary of the linguistic study that has been published in (Bouffier and Poibeau, 2007), since our aim in this paper is to show the machine learning process based on this linguistic analysis). These cues vary in nature: they can be based either on the material structure or the content of texts. We chose to mainly focus on task-independent knowledge so that the method is portable, as far as possible (we took inspiration from Halliday and Matthiessen's introduction to functional grammar (Halliday and Matthiessen, 2004). Some of these cues (especially connectors and lexical cues) can be automatically captured by machine learning methods.

- **Material structure cues.** These features include the recognition of titles, section, enumerations and paragraphs.
- **Morpho-syntactic cues.** Recommendations are not expressed in the same way as conditions from a morpho-syntactic point of view. We take the following features into account: 1) **Part of speech tags.** For example *recommandé* should be a verb and not a noun, even if the form is ambiguous in French; 2) **Tense and mood of the verb.** Present and future tenses are relevant, as well as imperative and conditional moods. Imperative and future always have an injunctive value in the texts. Injunctive verbs (see lexical cues) lose their injunctive property when used in a past tense.
- **Anaphoric cues.** A basic and local analysis of anaphoric elements is performed. We es-

pecially focused on expressions such as *dans ce cas*, *dans les N cas précédents*—(*in this case*, *in the n preceding cases...*) which are very frequent in clinical documents. The recognition of such expressions is based on a limited set of possible nouns that occurred in context, together with specific constraints (use of demonstrative pronouns, etc).

- **Conjunctive cues (discourse connectors).** Conditions are mainly expressed through conjunctive cues. The following forms are especially interesting: forms prototypically expressing conditions (*si*, *en cas de*, *dans le cas où...*—*if*, *in case of...*); Forms expressing the locations of some elements (*chez*, *en présence de...*—*in presence of...*); Forms expressing a temporal frame (*lorsque*, *au moment où*, *avant de...*—*when*, *before...*)

- **Lexical cues.** Recommendations are mainly expressed through lexical cues. We have observed forms prototypically expressing recommendations (*recommander*, *prescrire*,...—*recommend*, *prescribe*), obligations (*devoir*, ... *shall*) or options (*pouvoir*, ... *can*). Most of these forms are highly ambiguous but can be automatically acquired from an annotated corpus. Some expressions from the medical domain can be automatically extracted using a terminology extractor.

All these features are then formalized and translated into binary functions.

### 3.3 Formalisation

The lingusitic features described in the previous section vary in nature: they can be based either on the material structure or on the content of texts. The observations can be related to principles of Optimality Theory (McCarthy, 2008; Prince and Smolensky, 2004):

- Linguistic decision is relative, not absolute. Perfect satisfaction of all linguistic constraints is attained rarely, and perhaps never.

- Linguistic decision is a matter of comparison or competition among candidate output forms (none of which is perfect).

- Linguistic constraints are ranked and violable. Higher ranking constraints can compel violation of lower ranking constraints. Violation is minimal, however. And even low ranking constraints can make crucial decisions about the winning output candidate.

- The grammar of a language is a ranking of constraints. Ranking may differ from language to language, even if the constraints do not.

Even if our framework is different, these principles seem to be highly relevant for our task.

We chose to mainly focus on task-independent knowledge so that the method is portable, as far as possible (we took inspiration from Halliday and Matthiessen's introduction to functional grammar (Halliday and Matthiessen, 2004)). Some of these cues (especially connectors and lexical cues) can be automatically captured by machine learning methods. All these features are then formalized and translated into binary functions.

*\$Gov* (governor) refers to a condition governing a set of items; *\$Tar* (target) refers to a governed elements (either a recommendation or another condition).

**Cohesion cues**

- $\$Gov_i$ is syntactically integrated to the sentence; ⇒ *\$Gov* is integrated to the text if it does not appear at the beginning of a paragraph.
- $\$Gov_i$ and $\$Tar_i$ are in the same paragraph;
- $\$Gov_i$ contains a subordination cue;
- $\$Gov_i$ and $\$Tar_i$ are co-referring;
- $\$Tar_i$ includes a linguistic term expanding a term of $\$Gov_i$; ⇒ like *diarrhée* vs *diarrhée colique*.
- $\$Gov_i$ is similar to the section title; ⇒ i.e. contains the same plain words or a similar chunk of text.

**Breaking cues**

- $\$Gov_i$ is syntactically detached; ⇒ cf. *supra*.
- $\$Tar_i$ does not appear in the same paragraph as $\$Gov_i$;
- $\$Tar_i$ does not have the same layout as $\$Gov_i$;
- $\$Tar_i$ has the same lexical trigger as $\$Gov_i$;
- $\$Tar_i$ include a term with an extension of a term already in $\$Gov_i$;
- $\$Tar_i$ includes an antonymy cue with $\$Gov_i$;
- $\$Tar_i$ includes a coordination cue;
- $\$Gov_i$ is already linked to another recommendation.

We give below the set of constraints modelled from the lingusitic analysis above.

**Linguistic features modelled as binary functions**

- `is_syntactically_integrated($Gov)`
- `is_syntactically_integrated($Tar)`
- `are_in_the_same_paragraph($Gov, $Tar)`
- `are_in_the_same_visual_position($Gov, $Tar)` ⇒ for example, both appear at the beginning of a section
- `have_the_same_typographical_layout($Gov, $Tar)`
- `is_linked_with_the_title($Tar)`
- `located_after_a_subordination_connector ($Tar)`
- `located_after_a_coordination_connector( $Tar)`
- `have_same_lexical_trigger($Gov, $Tar)`
- `include_the_extension_of_a_term($Gov, $Tar)`
- `include_a_term_with_same_extension($Gov, $Tar)`
- `include_an_antonym($Gov, $Tar)`
- `are_co-referring($Gov, $Tar)`
- `is_already_linked_to_another_text_chunk ($Gov)`

Note that this part of the work requires deep linguistic analysis and deep linguistic knowledge in order to get accurate results. For example, co-reference is highly relevant in our case; thus, the system requires to get an accurate and operational co-reference solver adapted to the domain. However, it is now possible to find tools that are accurate enough to make the task tractable, even if some improvement in basic tools would of course improve the overall results.

### 3.4 Analysing the scope of conditions

Ultimately, our goal is to calculate the scope of conditions. We thus need to define the set of relevant features that will produce an *optimal* analysis of the scope of conditions (which means linking each pair of condition and recommendation accurately). Given the complexity of the task, it seems difficult to define manually a relevant set of rules, so we tried to automatically learn a classification function, using for training a set of pre-tagged examples.

We chose to represent the decision process through a decision tree. Decision trees provide an interesting framework for the task since they allow the expert to validate the results of the learning algorithm and possibly modify it. Moreover, it allows determining the relative weight and combinations of features. However, the data need first to be formalized in order to be used as input of the learning process.

Each constraint is represented as a binary feature: fourteen different features have been modelled as binary functions (see previous section). Each segment (condition or recommendation) is represented by a vector of 14 binary features (in other words, we get $2^{14} = 16384$ different possible instances).

The informativeness of each feature is then calculated using Information Gain (IG). Let $C$ be the set of constraints and $Ex$ the set of all training examples, *value(x,a)* with $x \in Ex$ defines the value of a specific example $x$ for attribute $a \in Attr$, $H$ specifies the entropy. IG is then defined as $KL(Ex,a) = H(Ex) - H(Ex|a)$.

The decision tree then integrates the information got from IG (i.e. the list of the most salient features and the most salient combination of features). In this experiment, we used the straightforward implementation (*J48*) provided with Weka.

## 4 Experiment

The approach described in the previous section has been evaluated on a set of French HPG.

### 4.1 Data

The training material is made of 18 HPG in French published by French national health agencies[2] between 2000 and 2005. These HPG focus on different pathologies (e.g. diabetes, high blood pressure, asthma etc.) as well as on clinical examination processes (e.g. digestive endoscopy).

### 4.2 Processing Steps

Segmenting a guideline to fill an XML template is a complex process involving several steps. We present briefly here the process that mainly implements what has been described in the previous section.

#### 4.2.1 Basic Segmentation

The text is first segmented in chunks corresponding to conditions and recommendations, following the description given in (Bouffier and

[2] ANAES, Agence Nationale d'Accréditation et d'Evaluation en Santé and AFSSAPS, Agence Francaise de Sécurité Sanitaire des Produits de Santé. Most of these practice guidelines are publicly available at: `http://www.anaes.fr` or `http://affsaps.sante.fr`. Similar documents have been published in English and other languages; the GEM DTD is language independent.

Poibeau, 2007). This is done automatically using linguistic cues; since the domain is specialized, the automatic analysis gives highly accurate results (P&R above .95[3]).

#### 4.2.2 Computing Frames and Scopes

The text is decomposed in pairs of basic segments (one recommendation with one condition at one time) as detailed in the previous section. Each pair is represented as a pair of constraint vectors. 800 manually annotated examples have then been provided for learning (400 positive exmaples and 400 negative). The system computes IG for each feature and then derives a decision tree.

New instances (new pairs of condition and recommendation) extracted from documents are then classified according to this decision tree. Since the scope of conditions is limited, we only take into account the ten conditions preceding a given recommendation (even if this threshold has been fixed manually, we observed that no recommendation is under the scope of a condition located before the tenth condition preceding the given recommendation). Among these ten candidiates, the system tries to find the optimal dependy tree.

## 5 Evaluation

We evaluated the approach on 5 HPG that have not been used for training, by measuring whether each recommendation is linked with the appropriate condition sequence or not.

### 5.1 Manual Annotation and Inter-annotator Agreement

The data is evaluated against HPG manually annotated by two annotators (a linguist and a domain experts). Inter-annotator agreement is high (*kappa* = 0.955), especially considering that we required an agreement between an expert and non-expert. This proves that the scope of conditions is expressed through linguistic cues which do not require, most of the time, domain-specific or expert knowledge. Yet the very few cases where the annotations were in disagreement were clearly due to a lack of domain knowledge by the non-expert.

[3] P&R is the harmonic mean of precision and recall (P&R = (2*P*R) / (P+R), corresponding to a F-measure with a $\beta$ factor equal to 1).

### 5.2 Evaluation of the Automatic Recognition of the Scope of Conditions

The scope of conditions is recognized with .78 accuracy — 531 out of 679 relations are accurately recognized— which is an interesting result given the large number of features involved in the process and the complexity of the rules.

To compare our results with a baseline, we reproduced the approach of (Georg and Jaulent, 2005), who perform the same task but limit their investigation to the sentence boundaries. Therefore, their approach cannot accurately calculate the scope of a condition if it spans several sentences. Our approach outperforms (Georg and Jaulent, 2005) by 30.65% (by 10 to 50% depending on the HPG), showing the necessity for a discourse-based approach that goes beyond sentence boundaries. The present experiment also outperforms the implementation of (Bouffier and Poibeau, 2007) by more than 5%, showing the interest of an automatic approach for the acquisition of rules.

Real-world experiments have been carried out with physicians who all said that the tool was useful in that it provides a first systematic analysis of HPG, even if manual work is of course required to obtain a fully accurate result. The tool can be used a step in a modelization process that involves several quality check examinations.

## 6 Related Work

GEM is an intermediate document model, between pure text (paper practice guidelines) and knowledge-based models like GLIF (Peleg et al., 2000) or EON (Tu and Musen, 2001). GEM is thus an elegant solution, independent from any theory or formalisms, but compliant with other frameworks like therapeutic algorithms.

GEM Cutter (`http://gem.med.yale.edu/`) is a tool aimed at aiding experts to fill the GEM DTD from texts. However, this software is only an interface allowing the end-user to perform the task through a time-consuming cut-and-paste process. The overall process described in (Shiffman et al., 2004) is also largely manual, even if it is an attempt to automate and regularize the translation process.

Several attempts have already been made to improve the use of practice guidelines: for example knowledge-based diagnostic aids can be derived from them (e.g. (Bouaud et al., 2001)). However, previous attempts to model automatically their meaning have been based on the analysis of isolated sentences and do not compute the exact scope of conditional sequences (Shiffman et al., 2000; Georg and Jaulent, 2005), with the exception of (Bouffier and Poibeau, 2007) who were using a limited set of manually defined rules for the task.

Our automatic approach relies on work done in the field of discourse processing. As we have seen in the introduction, the most important sequences of text to be tagged correspond to discourse structures (conditions, actions...). Although most researchers agree that a better understanding of text structure and text coherence could help extract knowledge, descriptive frameworks like the one developed by (Halliday and Hasan, 1976) are poorly formalized and difficult to apply in practice. D. Marcu was one of the first attempt to infer the rhetorical structure of a document using surface cues (Marcu, 2000). Since then, other techniques have been tried, with various results, as for example (Subba and Di Eugenio, 2007).

## 7 Conclusion

We have presented in this paper a system capable of performing an automatic segmentation of HPG. We have shown that our system outperforms previous systems: it is the first one capable of resolving the scope of conditions over several recommendations. In the future, we plan to apply our model to other languages and other kinds of texts. The task requires at least adapting the linguistic components of our system (mainly the preprocessing stage). More generally, the portability of discourse-based systems across languages remains a challenging area for the future.

## References

Jacques Bouaud, B. Séroussi, H. Dréau, H. Falcoff, C. Riou, M. Joubert, C. Simon, G. Simon, and A. Venot. 2001. ASTI : A Guideline-based Drug-ordering System for Primary Care. In *Medinfo 2001*, pages 528–532. IOS Press.

Amanda Bouffier and Thierry Poibeau. 2007. Automatic structuring of Practice Guidelines using the GEM DTD. In *BioNLP*. Prague.

Ross C. Brownson, Elizabeth A. Baker, Terry L. Leet, and Kathleen N. Gillespie. 2003. *Evidence-based Public Health*. Oxford University Press, Oxford.

Gersende Georg and Marie-Chrstine Jaulent. 2005. An Environment for Document Engineering of Clinical Guidelines. In *Medinfo 2005: Proceedings of the 10th World Congress on Medical Informatics*, pages 645–649. IOS Press.

Michael A.K. Halliday and Ruqaiya Hasan, editors. 1976. *Cohesion in English*. Longman, London.

Michael A.K. Halliday and Christian Matthiessen. 2004. *Introduction to functional grammar* . Arnold, London.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26:395–448.

John McCarthy. 2008. *Doing Optimality Theory*. Blackwell, Oxford.

Mor Peleg, A. Boxwala, O. Omolola, Qing Zeng, S. Tu, R. Lacson, E. Bernstam, N. Ash, P. Mork, L. Ohno-Machado, E.H. Shortliffe, and R.A. Greenes. 2000. GLIF3: The Evolution of a Guideline Representation Format. In *Medinfo 2001*, pages 645–649. IOS Press.

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell, Oxford.

Richard N. Shiffman, Bryant T. Karras, Abha Agrawal, Roland Chen, Luis Marenco, and Sujai Nath. 2000. Gem: A proposal for a more comprehensive guideline document model using xml. *Journal of the American Medical Informatics Assoc*, 7(5):488–498.

R.N. Shiffman, G. Michel, A. Essaihi, and E. Thornquist. 2004. Bridging the guideline implementation gap: a systematic, document-centered approach to guideline implementation. *Journal of the American Medical Informatics Assoc*, 11(5):418–426.

Rajen Subba and Barbara Di Eugenio. 2007. Automatic Discourse Segmentation Using Neural Networks. In *DECALOG*. Trento.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Articles*. PhD Thesis, University of Edinburgh.

Samson W. Tu and Mark A Musen. 2001. Modeling data and knowledge in the EON Guideline Architecture. In *Medinfo 2001*, pages 280–284. IOS Press.